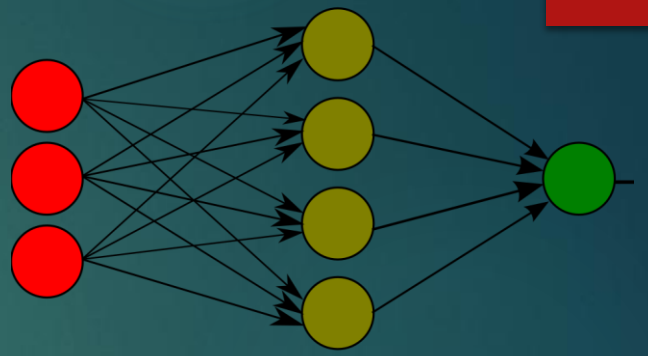
Come funziona e come sbaglia ChatGPT

FRANCESCO MORANDIN

Reti neurali artificiali

COME FUNZIONANO

Rete neurale artificiale



- Approssimatore parametrico

$$y = f(x, \theta) \text{ dove } f: \mathbb{R}^{n \times M} \to \mathbb{R}^m$$

- Addestramento: per imitazione sul dataset di training

$$(x_i, y_i)_{i=1,\dots,N}$$

- Inferenza: usato come predittore su nuovi dati

$$\hat{y}_j = f\left(x_j, \hat{\theta}\right)$$

- Si addestra minimizzando una loss iterativamente

$$\theta_0 \to \theta_1 \to \theta_2 \to \cdots \to \theta_T = \hat{\theta}$$

G.P.T.

- Generative Pre-trained Transformer
- Rete neurale che predice la prossima “parola”
- In realtà le parole sono spezzate in token

```
Per·ché· non· funz·iona· il· Wi··Fi·?
```

- L’input è la context window (da 2k a 1M token)
- L’output è la distribuzione di probabilità del prossimo token
- Si estrae a caso da questa distribuzione e si itera

Storia recente reti neurali

- Iniziano a funzionare dal 2012 (AlexNet, classificazione immagini)
- Più potenza e dati → più esperimenti → scoperte empiriche
- ReLU, data augmentation, normalization, residuals, Adam, attention, …
- Arrivano le LLM dal 2020 (GPT-3, 175B parametri)

Black Box

- Non c'è chiara comprensione teorica
- Molte scelte chiave sono evidenze empiriche
- Non sempre il training va a buon fine
- Tuttavia:

qualcosa nelle dimensioni, struttura dei dati e procedura di training le fa funzionare bene, meglio di quanto dovrebbero

Idee sbagliate comuni

QUANDO E PERCHÉ NON FUNZIONA

Stochastic Parrot

«Non ragiona, ma blatera a caso!»

- Idealmente non memorizza
- Sa generalizzare

  «When we train a large neural network to accurately predict the next word in lots of different texts from the Internet, what we are doing is that we are learning a world model. It may look on the surface that we are just learning statistical correlations in text. But it turns out that to 'just learn' the statistical correlations in text, to compress them really well, what the neural network learns is some representation of the process that produced the text. This text is actually a projection of the world.» (Ilya Sutskever, 2023)

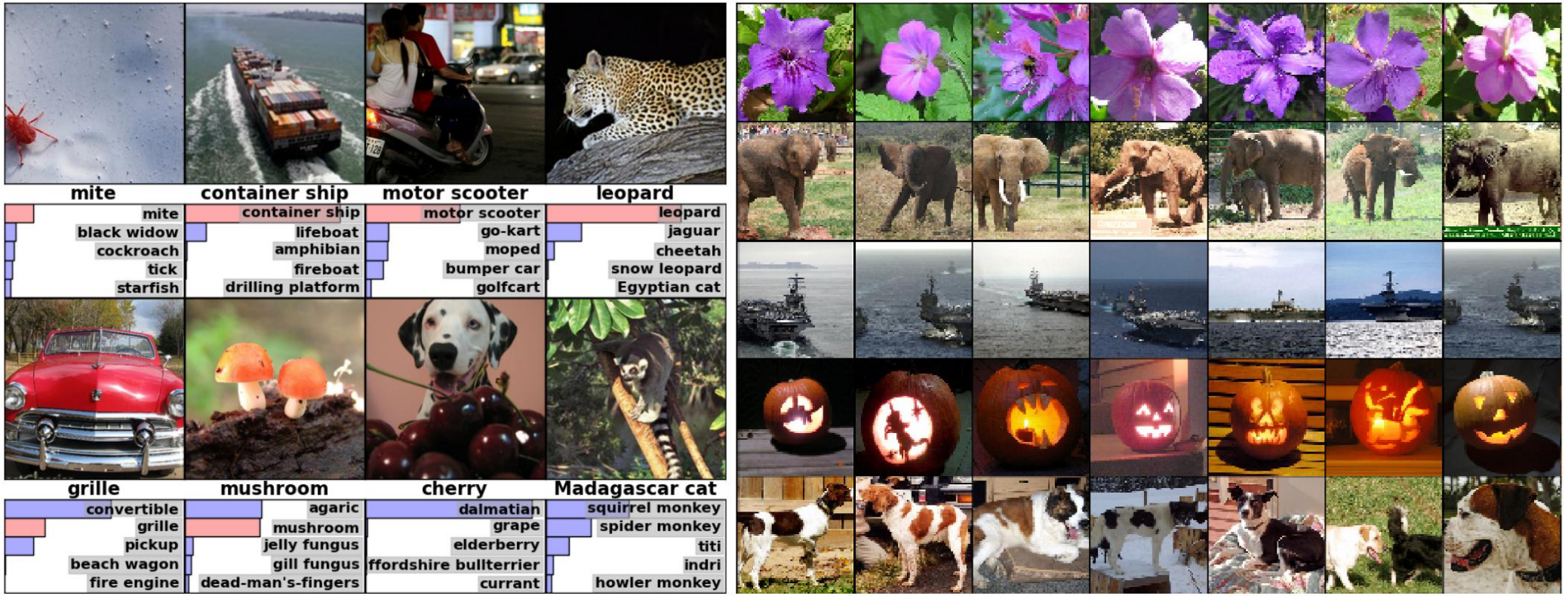
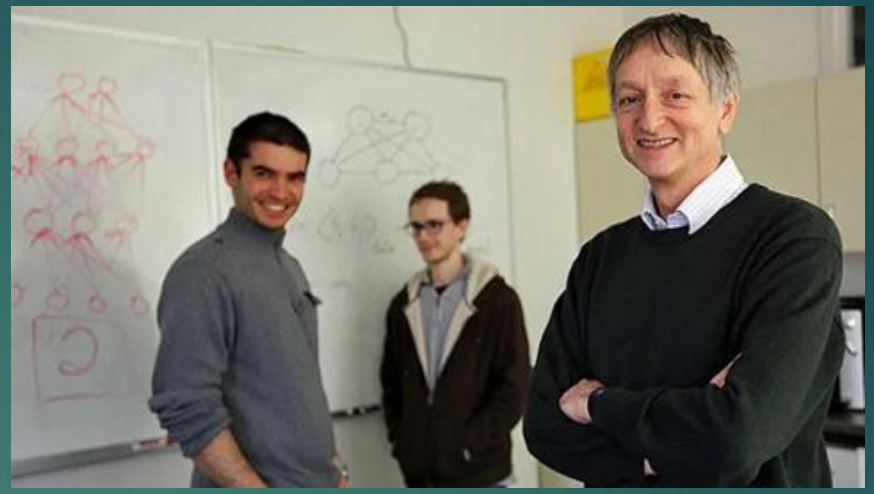


Figure 4: **(Left)** Eight ILSVRC-2010 test images and the five labels considered most probable by our model. The correct label is written under each image, and the probability assigned to the correct label is also shown with a red bar (if it happens to be in the top 5). **(Right)** Five ILSVRC-2010 test images in the first column. The remaining columns show the six training images that produce feature vectors in the last hidden layer with the smallest Euclidean distance from the feature vector for the test image.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. In *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*, 1097–1105.

Stochastic Parrot



- Idealmente non memorizza
- Sa generalizzare

«When we train a large neural network to accurately predict the next word in lots of different texts from the Internet, what we are doing is that we are learning a world model. It may look on the surface that we are just learning statistical correlations in text. But it turns out that to 'just learn' the statistical correlations in text, to compress them really well, what the neural network learns is some representation of the process that produced the text. This text is actually a projection of the world.» (Ilya Sutskever, 2023)

Garbage in, garbage out

«È addestrato su internet, non può essere affidabile!»

- Varie fasi di training
  - Pre-training su enormi quantità di testi da internet

    Impara sintassi, fatti, pattern, concetti astratti e un modello del mondo
  - Supervised fine-tuning

    Impara a rispondere da assistente, controllo di stile, specializzazioni di dominio
  - RL (HF e domini esatti)

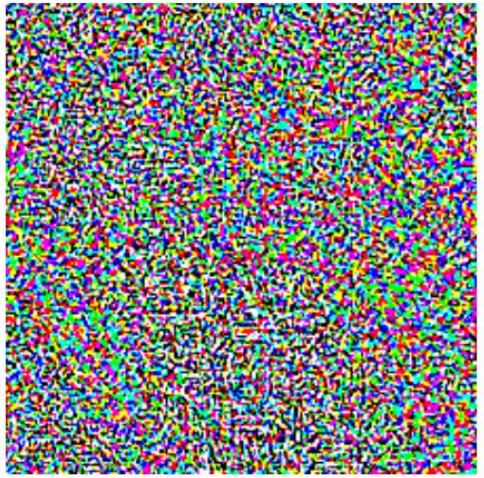
    Alza la qualità, la sicurezza, l'abilità di ragionare

Allucinazioni

«Mi ha dato una citazione che non esiste, inventa le cose!»

- I blind spot sono tipici della predizione e sono ben compresi
- Hanno a che fare con il comportamento microscopico della rete
- Esempio classico: adversarial attacks (images, go)
- Esistono approcci per aumentare la robustezza
- Esempio: flussi multipli (thinking)

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

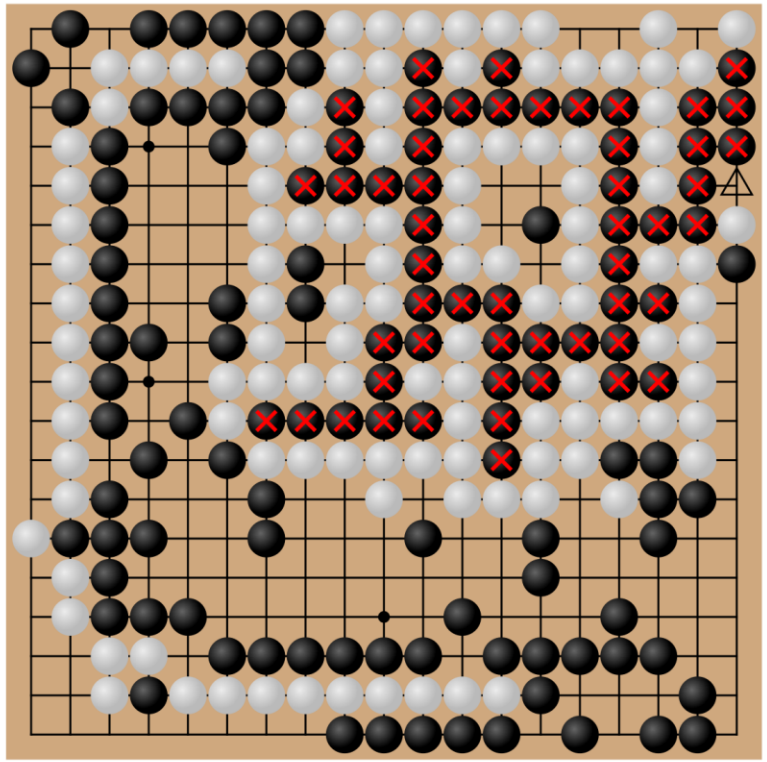
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

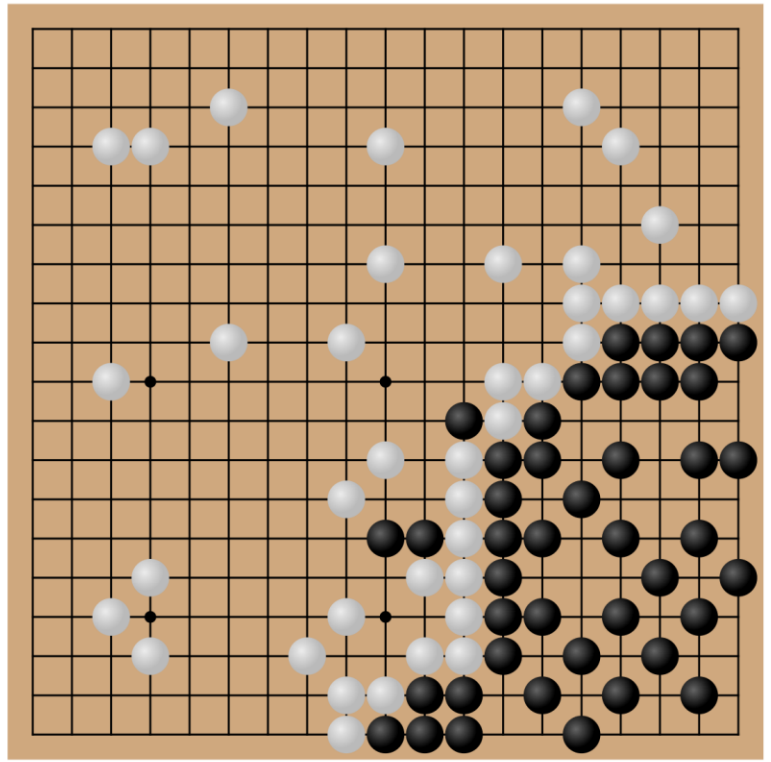
$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. In *International Conference on Learning Representations (ICLR)*.



(a) Our *cyclic-adversary* wins as white by capturing a cyclic group (×) that the victim (`Latest`, 10 million visits) leaves vulnerable. Explore the game.



(b) Our *pass-adversary* wins as black by tricking the victim (`Latest`, no search) into passing prematurely, ending the game. Explore the game.

Figure 1.1. Games between the strongest KataGo network at the time of conducting this research (which we refer to as `Latest`) and two different types of adversaries we trained. (a) Our *cyclic-adversary* beats KataGo even when KataGo plays with far more search than is needed to be superhuman. The adversary lures the victim into letting a large group of cyclic victim stones (×) get captured by the adversary's next move (Δ). Appendix J.2 has a detailed description of this adversary's behavior. (b) Our *pass-adversary* beats no-search KataGo by tricking it into passing. The adversary then passes in turn, ending the game with the adversary winning under the Tromp-Taylor ruleset for computer Go (Tromp, 2014) that KataGo was trained and configured to use (see Appendix A). The adversary gets points for its territory in the bottom-right corner (devoid of victim stones) whereas the victim does not get points for the territory in the top-left due to the presence of the adversary's stones.

Wang, T. T., Gleave, A., Tseng, T., Pelrine, K., Belrose, N., Miller, J., Dennis, M. D., Duan, Y., Pogrebniak, V., Levine, S., & Russell, S. (2023). Adversarial policies beat superhuman Go AIs. In *International Conference on Machine Learning (ICML)*.

Allucinazioni

- I blind spot sono tipici della predizione e sono ben compresi
- Hanno a che fare con il comportamento microscopico della rete
- Esempio classico: adversarial attacks (images, go)
- Esistono approcci per aumentare la robustezza
- Esempio: flussi multipli (thinking)

Cos'è Vilegis?

Vilegis è una piattaforma software italiana progettata per aiutare strutture ricettive (come hotel, agriturismi, B&B e campeggi) a gestire le proprie attività quotidiane in modo digitale e integrato.

🧩 Cosa fa Vilegis

Vilegis è un sistema gestionale "all-in-one" che centralizza diversi aspetti operativi:

- **Prenotazioni**: gestione di booking diretti e da portali (tipo OTA)
- **Channel manager**: sincronizza disponibilità e prezzi su più piattaforme (es. Booking, Airbnb)
- **Sito web con booking engine**: permette ai clienti di prenotare direttamente
- **Pagamenti online**: integrazione con sistemi di pagamento
- **CRM**: gestione dei clienti e comunicazioni
- **Marketing**: strumenti per offerte, newsletter e fidelizzazione

---

🎯 A chi è rivolto

È pensato per:

Cos'è Vilegis?

Controllo al volo cosa indica “Vilegis”, perché potrebbe essere un nome recente, di nicchia o scritto in modo specifico.

Thought for 11s >

“Vilegis” di solito si riferisce a **Battle for Vilegis**, un **gioco di ruolo dal vivo (LARP)** ambientato in un universo fantasy. Sul sito ufficiale viene descritto come un evento di **gioco di ruolo dal vivo** con più appuntamenti, e altre fonti lo presentano come uno dei principali eventi LARP in Italia. battleforvilegis.c... +2

In pratica, i partecipanti:

- interpretano un personaggio,
- prendono parte a battaglie, missioni e trame narrative,
- si inseriscono in diverse casate o fazioni del mondo di gioco. Il materiale ufficiale parla infatti di giocatori che giurano fedeltà a una delle casate combattenti del setting. battleforvilegis.c... +1

Quindi **Vilegis non è un software o un servizio comune**: è soprattutto il **nome del mondo/evento LARP** legato a *Battle for Vilegis*. battleforvilegis.c... +1

Illusione dell'oracolo

«Se fosse davvero intelligente, non fallirebbe in questo!»

- Le LLM non sono (ancora) persone, ma strumenti
- Comprensibili solo superficialmente e prive di manuale
- Punti di forza e debolezza sono diversi dai nostri
- Possono fallire se messi alla prova come intelligenze generali

  «Spiega la spooky action at a distance di Einstein»
- Possono essere messe in difficoltà involontariamente

  «Scrivi le soluzioni ai problemi di test2.tex usando lo stile di test1_sol.tex»
- Spesso non fallisce sul singolo passo, ma nella composizione di molti passi

Per un uso consapevole

- Usare i modelli “thinking”
- Imparare pregi e difetti del modello scelto
- Tenere presente che l’input è la context window
- Considerare il system prompt
- Ricordare che è in parte casuale
- Ogni step di complessità peggiora i risultati
- La lingua non è indifferente
- Illusione di rigore

Per un uso consapevole

- Tre tipi di uso
  - Uso linguistico-redazionale: riscrivere, riassumere, spiegare, produrre varianti, tradurre, generare esempi.

    Qui i modelli sono spesso molto forti
  - Uso concettuale-esplorativo: brainstorming, confronto tra definizioni, intuizioni, esempi e controesempi, possibili strategie di dimostrazione.

    Qui possono essere molto utili, ma vanno verificati
  - Uso formale-dimostrativo: teoremi, dimostrazioni, citazioni bibliografiche, calcoli delicati, riferimenti storici, passaggi tecnici sensibili.

    Qui serve controllo rigoroso, e il rischio di errore è alto